FEATURE SELECTION AND WEIGHING FOR CASE-BASED REASONING SYSTEM USING RANDOM FORESTS

BOOMA DEVI SEKAR, HUI WANG
School of Computing, Ulster University
Northern Ireland, UK

Case-based reasoning has become a successful technique that uses the previous experience as a problem-solving paradigm. It adapts or reuses the solutions of a similar problem to solve a new one. In a case-based reasoning system, it is important to have a good similarity retrieval algorithm to retrieve the most similar cases to the query case. However, we also note that in a medical domain with increased use of electronic health records, the availability of patient cases and the related attributes have increased. Thus, as a pre-processing step or as part of the retrieval algorithm, it becomes critical to select the most informative features to improve the retrieval efficiency and accuracy in a case-based reasoning system. In this paper we explore random forest, a popular method in machine learning, for feature selection and weighting in a case-based reasoning system and investigate the case retrieval accuracy.

1. Introduction

In recent years with the rapid increase in electronic health record (EHR) adoptation, there has been a growing expectation in the development of personalised medicine, which aims to customize treatment for an individual patient based on their likelihood of response to a therapy. The move towards personalized medicine is supported by various technological advancements, escpecially in the area of machine learning and artificial intelligence. One such pathway is the development of personalized diagnostic/ predictive model based on patient similarity.[1] Such model aims to identify and derive insights from patients similar to the query (new) patient and then analyze the derived insights in the diagnostic/ prediction model to provide personalized treatment/ prediction to the query patient.

Case-based Reasoning (CBR), an artificial intelligent approach has become the most trustworthy methodology for developing personalized diagnostic/ predictive model that is very close to human reasoning.[2] CBR adapts a supervised learning algorithm, which trains on previous experience in form of resolved cases stored in the case base to provide a solution to the new problem. Thus, unlike various other AI approaches such as rule-based reasoning, or neural networks, that generate abstract representations from a set of training examples, CBR methodology adapts instance-based learning and uses previous similar cases as

the basis for decision making. A generic CBR system is composed of four consecutive processes, known as the CBR cycle, including retrieval, reuse, revise and retain.[3] The first and most important step is 'retrieval' that applies similarity retrieval algorithm in search for the most similar cases from the case base. The subsequent step, 'adaptation' (reuse and revise) uses the information acquired from the retrieved case(s) to solve the query case. Finally, 'retain' learns from the problem-solving experience and stores the new knowledge in the case base for solving future new problems. Among these four steps, case retrieval efficiency and accuracy are topics of great interest among researchers. For which many researchers focused on improving the retrieval performance by developing different similarity measure algorithms.[4,5] However, an important step overlooked, that could improve the case retrieval performance is the selection of the informative features for CBR system.

Moreover, with the increased use of EHR and big data in healthcare, availability of a large number of patient cases and the relevant clinical attributes are becoming more common. Although such large case base and clinical attributes could increase the coverage of the application domain to provide a solution for the new query case, it does increase the possibility of having irrelevant and redundant features in the case base, which in turn affects the retrieval performance. Thus, as a pre-processing step or as part of the model, it becomes critical to select the most informative features for building any diagnostic/ predictive model. In CBR system, feature selection and weighting could determine the representative features required and remove the redundant ones.

In this paper, we investigate Random Forests (RF)[6] algorithm for feature selection and weighting for a CBR system. The contribution of the paper is the comparative assessment of the univariate, recursive feature elimination (RFE), RFE with cross validation (RFE-CV) and tree based feature selection methods with RF to compute the feature weighting for a CBR system. The main goal is to examine on whether the selected important feature variable using the above methods could improve the retrieval efficiency and accuracy of a CBR system.

In the following section, we present the technical background of the feature selection method and similarity retrieval measure applied. In Sec. 3, we evaluate the feature section method by analyzing the breast cancer database obtained from Breast Cancer Surveillance Consortium (BCSC). Finally, in Sec. 4, we present the conclusion drawn from the results obtained and propose future research work in such research area.

2. Background

2.1. *Feature Selection with Random Forest*

In general, feature selection can be categorized as a filter, wrapper, and embedded methods. The filter method, execute feature selection independent of the chosen predictor model. It treats feature selection as a pre-processing step, and are usually

applied to remove some spurious features from the data set and are not much useful for measuring the feature importance. The wrapper, on the other hand, estimates the feature importance of variables by evaluating the model of interest. They are generally based on a black box evaluator and therefore are constrained by the given predictor model and search strategy. Finally, the embedded method performs feature selection as a part of the model building process and are generally found to be more beneficial, but face the challenge of over-fitting.

In order to combine the best properties of the above three methods, in this paper we explore hybrid feature selection methods. Study shows that with different combinations, hybrid methods could achieve higher efficiency and accuracy in feature selection. In this paper, we will investigate univariate chi-square[7], recursive feature elimination (RFE)[8] and tree based feature selection[9] with RF for feature weighting in a CBR system.

RF is a non-parametric and highly flexible model and thus when applied for measuring variable importance, it could capture both linear and non-linear relations in the data. It has an embedded feature ranking technique: 'variable importance measure', which can be used as a tool to select the important features aiding the predictor model. A simple variable importance measure would count the number of times each variable is selected by all individual trees in the ensemble. In this paper, a Gini index, which measures the weighted mean of the individual trees improvement at the splitting point is used to measure the variable importance. Gini index based measure can be computed in RF using Eq. (1).

$$G(t) = 1 - \sum_{k=1}^{Q} p^2(k|t) \quad (1)$$

Where, for a given node t and estimated class probabilities $p(k|t)$, $k = 1, \ldots, Q$ and Q is the number of classes.

2.2. *Similarity Retrieval Measure for Case Based Reasoning System*

In the proposed model, the similarity retrieval algorithm is defined using "local" and "global" similarity functions. Local similarity function measures the distance between the simple attributes, whereas the global similarity function applies the results from local similarity measures to compare the compound attributes. In the proposed CBR system, a patient case is represented as a compound attribute, composed of several simple attributes, including physiological and clinical variables. Thus, local similarity functions are first applied to compute the distance between simple attributes in the query case against the ones characterizing patient cases in the case base. The result of local similarity measures of all simple attributes is then aggregated using the global similarity function to select the patient case(s) that are most similar to the query case from the precedent ones present in the patient case base. K-Nearest Neighbour (k-NN) is computed as the global similarity function to retrieve the top k similar cases to the query case from the patient case base. Given the query patient case $X = x_i$ and the local similarity measure computed for the N

number of patient cases in the case base $Y_j = y_{ji}$ the Euclidian distance between the query and case base is computed using Eq. (2).

$$d(X, Y_j) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_{ji}^2} \quad (2)$$

Based on the above result, the k nearest patient cases are first located in the case base. The k-NN similarity measure is then computed, which measures the arithmetic mean output across patient cases in the case base and returns a value between 0 ~ 1, with 0 and 1 indicating the retrieved case being less and most similar to the query case, respectively.

3. Evaluation

For evaluating the feature selection using RF, we analyse the breast cancer database obtained from BCSC. The database consists of 2,392,998 index-screening mammograms from women who were not diagnosed with breast cancer previously. In order to have a manageable size of the dataset, BCSC performed a cross-classification of risk factors and outcome and aggregated the patient cases based on the frequency of each combination. The reduced dataset, now with categorical variables consists of 280,660 records. Among which, 6274 were diagnosed with invasive or ductal carcinoma in situ in breast and 175629 with no cancer. Various research works have been conducted based on BCSC database, some of which include, pathology identification,[10] examining patterns in mammography for different ethnic group,[11] and genetic testing for breast cancer.[12] Table 1 shows the breast cancer patient variables and corresponding categorization (coding) defined in the dataset. Here the variables 1-12 were used as the input variables and variable 13 as the classification variable for the analysis. The dataset was split into 70% for training and 30% for testing.

Table 1. List of breast cancer patient variables in BCSC risk-estimate database.

Variable	Name	Coding
1	menopause	0 = premenopausal; 1=postmenopausal; 9 = unknown
2	agegrp	1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84
3	density	BI-RADS codes: 1 = almost entirely fat; 2 = scattered fibroglandular densities; 3 = heterogeneously dense; 4 = extremely dense; 9 = unknown
4	race	1 = white; 2 = Asian/ Pacific islander; 3 = black; 4 = Native American; 5 = other/mixed; 9 = unknown
5	hispanic	0 = no; 1 = yes; 9 = unknown
6	bmi	Body mass index: 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = unknown
7	agefirst	Age at first birth: 0 = age < 30; 1 = age ≥ 30; 2 = Nulliparous; 9 = unknown
8	nrelbc	Number of first degree relatives with breast cancer: 0 = zero; 1 = one; 2 = 2 or more; 9 = unknown
9	brstproc	Previous breast procedure: 0 = no; 1 = yes; 9 = unknown

10	lastmamm	Result of last mammogram before the index mammogram: 0 = negative; 1 = false positive; 9 = unknown
11	surgmeno	Surgical menopause: 0 = natural; 1 = surgical; 9 = unknown or not menopausal
12	hrt	Current hormone therapy: 0 = no; 1 = yes; 9 = unknown
13	cancer	Diagnosis of invasive or ductal carcinoma in situ breast cancer within one year of the index screening mammogram: 0 = no; 1 = yes

Evaluating with RF, four tests were performed on the dataset. In the first test, no features were selected and thus all variables (1-12) in Table 1 were used to classify the patients with no cancer and invasive cancer or ductal carcinoma in situ breast cancer. In the second test, univariate feature selection using chi-square method with a number of features to be extracted 'k' = 7 was applied. This extracted the 7 best features (agegrp, race, brstproc, nrelbc, bmi, density, and surgmeno), which were then used to classify the patients using RF. In the third test – RFE is applied by pre-specifying the number of features to be extracted 'k' = 7. In this method, weights are initially assigned to each feature, it then recursively eliminate the features whose absolute weight is smallest and extracts the 'k' most important features. The 7 best features extracted using RFE were agegrp, density, race, bmi, agefirst, nrelbc and hrt. We note that, compared with the univariate feature selection method, RFE method extracted two different features, namely agefirst and hrt.

In the last test, RFE-cross validation (RFE-CV) was applied. With cross validation, the optimal number of features to obtain the best accuracy and the relevant important features were identified using RFE. Three optimal features, namely agegrp, density, and bmi were extracted using the RFE-CV method. Finally, using tree based feature selection with RF, the feature importance method was applied to eliminate the correlated feature in each iteration and list all the attributes according to its feature importance in solving the classification problem. Table 2 presents the classification results of the above four tests in terms of accuracy, sensitivity, specificity, true positive rate, true negative rate, false positive rate and false negative rate. Fig. 1 shows the sequence of feature importance of the input attributes using tree-based feature importance measure using RF.

Table 2. Statistical results of the four tests presented in Figure 1.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	True Positive Rate	True Negative Rate	False Positive Rate	False Negative Rate
a	94.76	6.11	97.89	0.093	0.9671	0.021	0.9388
b	96.49	0	99.91	0	0.9657	0.0008	1
c	96.51	0	99.93	0.026	0.9658	0.0008	0.9994
d	96.50	0	99.92	0	0.9657	0.0008	1

Method: (a) RF with no feature selection (b) Univariate feature selection using Chi2, K = 7 and RF (c) RFE-RF with k=7 (d) REF-CV and RF: Optimal number of features = 3

The results in Table 2 show that a better accuracy is achieved when feature selection method was applied. Having comparatively less number of cases (6274) with invasive cancer or ductal carcinoma in situ in the breast than the ones with no cancer (175629), the classification result show a poor sensitivity but a good specificity. The same can be observed for the true negative rate (closer to 1 is better) and false positive rate (closer to 0 is better) when compared to the true positive rate (closer to 1 is better) and false negative rate (closer to 0 is better).

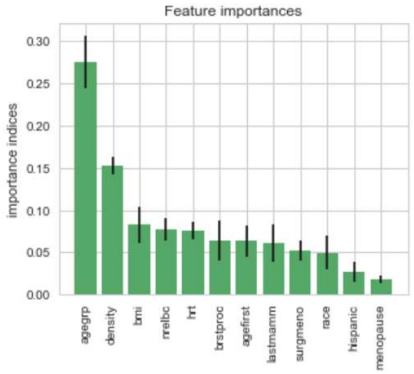

Figure 1. Variables listed according to the feature importance using RF.

As the main purpose of this paper is to evaluate on whether the feature importance using RF is useful for assigning weights to the description variables in the CBR system. As a case study, the features extracted using RF were applied to CBR system to evaluate on whether similar patient cases could be retrieved. Table 3 shows the results of similar cases retrieved for the same query case. Assigning $k = 3$ in the k-NN similarity measure, three similar cases were retrieved for each of the conditions presented above. Two tests were performed to evaluate on whether reducing the number of description variables in a CBR system could affect the performance of the case retrieval. In both the tests, wt = 1 was assigned to the important attributes, meanwhile, in the first test, the non-important attributes were eliminated by assigning wt = 0 and in the second test, a minimum wt = 0.1 was assigned to still include them in the computation of k-NN similarity measure.

Table 3 show that in both the tests better accuracy was achieved when feature selection method was applied for feature weighting in the CBR system. However, when observing the attribute values which are different from the query case, we note that in the second test, in predominant of the cases, only one attribute was different from the query case. Whereas, in the first test, for the cases retrieved, many times two attributes were different from the query case, showing that it could not retrieve cases which were most similar to the query case.

Table 3. Case retrieval results for a query case in a CBR system using the results of RF.

	Wt = 1 assigned to important attributes Wt = 0 is assigned to all other attributes			Wt = 1 is assigned to important attributes Wt = 0.1 is assigned to all other attributes		
	Case id	Retrieval accuracy (k-NN)	Attributes different from query case	Case id	Retrieval accuracy (k-NN)	Attributes different from query case
Query Patient	10000			10000		
Equal weight to all attributes (Wt = 1 assigned for all attributes)	10003	0.9166	agefirst	10003	0.9166	agefirst
	10495	0.9166	race	10495	0.9166	race
	10496	0.9166	race	10496	0.9166	race
Univariate - Chisquare, SelectKBest (K=7)	10003	1	agefirst	10003	0.9866	agefirst
	10002	1	agefirst, lastmamm	9999	0.9866	lastmamm
	10001	1	agefirst, nrelbc	9947	0.9866	hispanic
RFE with RF (K = 7))	9883	1	hispanic	9999	0.9866	lastmamm
	9947	1	hispanic	9947	0.9866	hispanic
	9946	1	hispanic, lastmamm	174378	0.973	hispanic, lastmamm
RFE with RF (with 3 Optimal features)	10001	1	agefirst, nrelbc	10003	0.9743	agefirst
	10002	1	agefirst, lastmamm	10495	0.9743	race
	10003	1	agefirst	10496	0.9743	race
RFE – CV (using distributed weight assigned according to variable importance measure with RF)	9947	0.981	hispanic	9947	0.981	hispanic
	10003	0.944	agefirst	10003	0.944	agefirst
	10495	0.944	race	10495	0.944	race

4. Conclusion and Future Work

RF is a popular machine learning tool, frequently applied in solving various scientific problems, from feature selection, regression to classification. In this paper, we investigate RF for feature weighting the description variables in the CBR system and examine on whether feature selection could improve the case retrieval performance. Through evaluating the hybrid feature selection methods, including univariate, RFE, and tree based feature selection with RF on a breast cancer database obtained from BCSC, we conclude that feature selection with RF is a sensible approach for feature weighting in a CBR system. However, feature selection has to be done in two stages, first for a large dataset, all the spurious features have to be eliminated. This can be done by applying a filter method, such

as the chi-square method, tested in this paper. Secondly, the feature importance method using RF or RFE-CV can be applied to select the important features. In terms of feature weighting for CBR system, to improve the case retrieval performance, it would be sensible to either use distributed weighting or assign minimum weight to the non-important attributes.

For future work, it would be worthwhile to test other feature selection methods such as Pearson correlation coefficient, mutual information, Gram-Schmidt orthogonalization and hybrid feature selection methods. Also, evaluate on how they perform in comparison with the methods using RF presented for feature weighting in a CBR system. To assess the impact of feature selection methods, it would be useful to evaluate the performance of the feature selection methods using error rate and time scores performance matrices.

Acknowledgments

The DESIREE project has received funding from the European Union´s Horizon 2020 research and innovation program under grant agreement No. 690238.

References

1. C.L. Parra-Calderón, Patient Similarity in Prediction Models Based on Health Data: A Scoping Review, *JMIR Med Inform.* 5(1), (2017).
2. X. Blanco, S. Rodríguez, J.M. Corchado, C. Zato, Case-based Reasoning applied to Medical Diagnosis and Treatment, *AISC,* 217 (2017).
3. J. Chen, Z. Teng, Z. Liu, A Review and Analysis of Case-based Reasoning Research, *ICITBS,* (2015).
4. D. Shasha, WALRUS: A Similarity Retrieval Algorithm for Image Databases, *IEEE TKDE,* (1999).
5. L. Liu, Z. Lin, L. Shao, Sequential Discrete Hashing for Scalable Cross-Modality Similarity Retrieval, *IEEE TIP,* 26(1), (2017).
6. L. Breiman, Random Forests, *Machine Learning,* 45, 5-32, (2001).
7. A.M. Bidgoli, M.N. Parsa, A Hybrid Feature Selection by Resampling, Chi-squred and Consistency evaluation Technique, IJCIE, 6(8), (2012).
8. X. Zeng et al., D.V. Alphen, Feature Selection using Recursive feature Elimination for Handwritten Digit Recognition, *IIH-MSP'09,* (2009).
9. G. Louppe et al., Understanding variable importance in forests of randomized trees, *Adv in Neural Information Processing Sys,* (2013).
10. D.L. Weaver et al., Pathologic findings from the Breast Cancer Surveillance Consortium, *Cancer,* 106(4), (2006).
11. F.D. Gilliland et al., Patterns of mammography using among Hispanic, American Indian and non-Hispanic White women in New Mexico, *Am J Epidemiol,* 152(5), (2000).
12. C.M. Velicer, Genetic testing for breast cancer: where are health care providers in the decision process, *Genet Med,* 3(2), (2001).